A.V. Smirnov, T.V. Levashova, M.P. Pashkin, N.G. Shilov,
A.A. Krizhanovsky, A.M. Kashevnik, A.S. Komarova

CONTEXT-SENSITIVE ACCESS TO E-DOCUMENT CORPUS[1]

1. Introduction

The methodology of context-sensitive access to e-documents considers context as a problem model based on the knowledge extracted from the application domain, and presented in the form of application ontology.

Efficient access to an information in the text form is needed. Wiki resources as a modern text format provides huge number of text in a semi formalized structure.

At the first stage of the methodology, documents are indexed against the ontology representing macro-situation. The indexing method uses a topic tree as a middle layer between documents and the application ontology. At the second stage documents relevant to the current situation (the abstract and operational contexts) are identified and sorted by degree of relevance. Abstract context is a problem-oriented ontology-based model. Operational context is an instantiation of the abstract context with data provided by the information sources.

The following parts of the methodology are described: (i) metrics for measuring similarity of e-documents to ontology, (ii) a document index storing results of indexing of e-documents against the ontology;

[1] This work was partly supported through project № 16.2.35 of the research program «Mathematical Modelling and Intelligent Systems», project № 1.9 of the research program «Fundamental Basics of Information Technologies and Computer Systems» of the Russian Academy of Sciences, the project funded by grant № 05-01-00151 of the Russian Foundation for Basic Research. The presented research was partially supported through CRDF partner project № RUM2-1554-ST-05 with US ONR and US AFRL.

(iii) a method for identification of relevant e-documents based on semantic similarity measures.

Wikipedia (wiki resource) is used as a corpus of e-documents for approach evaluation in a case study. Text categorization, the presence of metadata, and an existence of a lot of articles related to different topics characterize the corpus.

3. Approach

Document Index stores results of indexing the e-documents against the application ontology. One of the indexing purposes is to calculate and store the similarity of a document to a fragment[1] of the application ontology. The approach is described in our previous work[2] in detail.

The Index contains a set of tuples $<A^{II}, Doc\_ID, Sim>$, where A^{II} – application ontology fragment to which the document is similar; *Doc_ID* – document identifier; and *Sim* – value of similarity of the document to the application ontology fragment. The initial set of ontology classes (for the document) is selected by comparing text similarity between class names and document's text / metadata.

3.1. Topic based index creation

The index is created between Wikipedia documents and ontology classes and attributes via Wikipedia categories. The algorithm has the following parameters:

$Wikipedia_{Articles}$ – set of Wikipedia articles.

ONT – set of names of classes and attributes from ontology.

D_{max} – maximum distance between similar words.

---

[1] Ontology fragment contains sets of classes with relationships between them, class attributes, and domains for string attributes.

[2] *Smirnov A., Levashova T., Shilov N., Pashkin M., Kashevnik A., Krizhanovsky A., Komarova A.* Context-Sensitive Access to Information Sources // Proceedings: International Conference on Hybrid Information Technology. Cheju Island, Korea, November 9th-11th, 2006. Accepted for publication.

O_{res} – result set of classes and attributes from ontology that are correspond to the document.

k – weight coefficient in [0,1].

$|CS_{max}|$ - number of categories in Wikipedia.

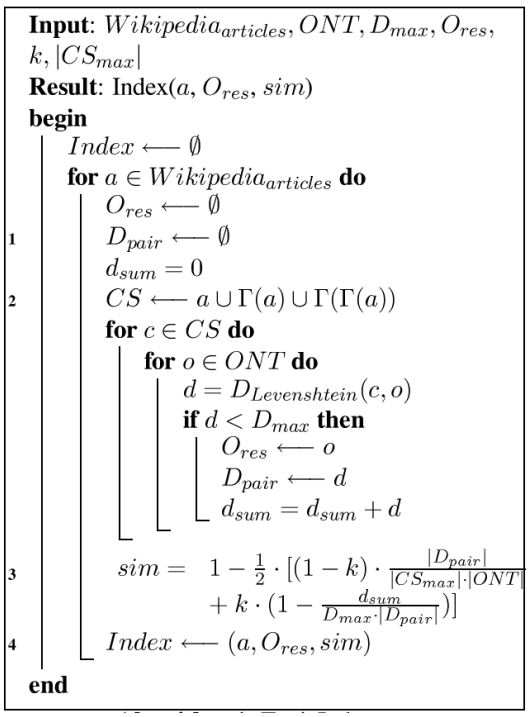

Fig. 1. Topic-based indexing algorithm

Algorithm of topic-based index creation is designed and implemented. The short description of following steps is provided:

1) D_{pair} is the array of Levenshtein distances between (1) names of ontology classes, attributes and (2) titles of articles, categories of Wikipedia.

2) The set of neighbours categories of categories of article *a* is added to *CS*.
3) The value of *sim* (step 11) is in the range [0,1], because:

- weight coefficient *k*·is in the range [0,1];
- $0 \leq |D_{pair}| / (|CS_{max}| \cdot |ONT|) \leq 1$ (size of subset of similar pairs). The size of CS_{max} is a constant (number of categories in Wikipedia) here, but it is enough to require that CS_{max} is not less than the maximum size of the set |CS| for each article from Wikipedia $_{Articles}$ in order to satisfy this inequality. But it will require some additional computations;
- $0 \leq d_{sum} / (D_{max} \cdot |D_{pair}|) \leq 1$ (ratio of Levenshtein distance sum to maximum possible distance from document text / metadata to ontology found element names O_{res});
- let's sim:=1 if the array D_{pair} (pairs of similar names of ontology classes and Wikipedia categories) is empty.

4) The set of found ontology classes and attributes (that have names similar to the title of Wikipedia article or categories of the article) is stored to Index.

The value of *sim* 1 means high similarity of set of ontology classes and attributes to the document, 0 – similarity is absent.

3.2. Relevance estimation

At the second stage of approach, when Index $<A^{II}, Doc\_ID, Sim>$ was already created, the documents were evaluated by their relevance to the current situation. Measurement of similarity between contexts (as ontology model) and documents is based on a comparative assessment of these contexts and application ontology fragments A^{II}, elements of which have been included in the contexts.

In graph-based method the fragment A^{II} and abstract context are represented as two graphs. The graphs are compared taking into account (1) similarity between names of graph nodes; (2) number of

neighbouring nodes for the similar nodes; (3) the shortest paths between similar nodes.

The fragments are sorted according to their similarity to the abstract context. The fragments having greater similarity are considered as more relevant in this particular situation. Since for each document there is a fragment corresponding to it, documents are sorted by their relevance to the abstract context as well.

5. Conclusion

Context-sensitive access to e-documents approach is described in the paper. The ontology fragment selection algorithm is presented briefly. The structure of index (documents are indexed against the ontology) and topic-based indexing algorithm are described. The graph-based method for identification of relevant e-documents based on semantic similarity measures is described.

The logical evolution of the current approach is a clustering of similar documents for search improving. It is possible, since similarities of documents against fragments of problem-oriented ontologies are calculated. Thus, a function aggregating these similarities as a basis for documents clustering should be developed.

The document classification could be improved by using the linguistics resources such as, Eurovoc and WordNet. At the same time, the possibility to cluster documents should allow classifying new documents, since (1) Wikipedia document classified via Wikipedia categories, (2) Wikipedia documents and new (non Wikipedia) documents will share the same cluster, due to huge number of Wikipedia articles in different languages for different problem-oriented domains.